Joint BecA Hub and UNESCO Advanced Genomics and Bioinformatics: Viral/Bacterial Metagenomics and Next generation sequencing Workshop

Introduction to Bioinformatics Course

13th - 17th August 2012

ILRI-BecA, Nairobi

Etienne de Villiers
International Livestock Research Institute
Nairobi, Kenya

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

EMBnet
The Global Bioinformatics Network

Acknowledgements

This course was adapted from a course designed and implemented by David Lynn and Andrew Lloyd while working at the Education and Research Centre (ERC) at St. Vincent's University Hospital, Dublin. The original course and manual implemented by David Lynn grew naturally from The ABC Bioinformatics Course, an earlier Irish National Centre for BioInformatics (INCBI) project based on GCG and the WWW, to which Aoife McLysaght (TCD) was a major contributor. That in turn owes a debt of gratitude to the ABCT tutorial designed by Rodrigo Lopez when he was the Norwegian EMBnet node. This course would never have got off the ground without the encouragement of Cliona O'Farrelly, the Research Director at the Education and Research Centre (ERC) at St. Vincent's University Hospital. The development of the original course was funded by the Dublin Molecular Medicine Centre and the Conway Institute, University College Dublin.

The Multiple Alignment and Phylogenetics section were adapted from a course developed by Hans-Henrik and Anders Gorm Pedersen, SLU, Sweden.
Ensembl section was adapted from Ensembl tutorials and worked examples on Ensembl website.

Table of Contents

Introduction 4

Introduction to Bioinformatics 6

Databases 6

Sequence formats 11

Accession numbers 12

Interrogating (sequence) databases 14

Ensembl – http://www.ensembl.org 14

SRS - http://srs.ebi.ac.uk/ 18

Entrez - http://www.ncbi.nlm.nih.gov/Entrez/ 24

Nucleic Acid Sequence Analysis 25

1) Nucleic acids and the genetic code 25

2) Translating DNA in 6-frames: 27

3) Reverse Complement & other tools: 29

4) Oligo Calculator - http://www.pitt.edu/~rsup/OligoCalc.html 31

Protein Sequence Analysis 33

1) Physico-chemical properties: 33

2) Cellular localization: 35

3) Signal peptides: 37

4) Transmembrane domains: 39

5) Post-translational modifications: 42

6) Motifs and Domains 44

7) Secondary Structure Prediction 45

Printed sources about Bioinformatics and the Internet. 47

APPENDIX I 48

APPENDIX II 50

Introduction

This course is designed to impress upon you that computers and the Internet can not only make your work as a biologist easier and more productive but also enable you to answer questions that would be impossible without computational help. Thus there are some computational analyses that you could conceivably do on the back of an envelope or with a pocket calculator and there are others so computationally demanding that you would not attempt them without electronic help. An example of the first would be to scan the following DNA sequence for ecoRI restriction endonuclease sites (GAATTC):

```
>Adhr D.melanogaster
ATGTTCGATTTGACGGGCAAGCATGTCTGCTATGTGGCGGATTGCGGAGGGAGACCAGC
AAGGTTCTCATGACCAAGAATATAGCGAAACTGGCCATTCGGAAAATCCCCAGGCCATC
GCTCAGTTGCAGTCGATAAAGCCGAGTACTTCTGGACCTACGACGTGACCATGGCAAGA
ATTCATATGAAGAAGTACTGATGGTCCAAATGGACTACATCGATGTCCTGATCAATGGT
GCTACGCTGATAACATTGATGCCACCATCAATACAAATCTAACGGGAATGATGAACACG
TGTTACCCTATATGGACAGAAAAATAGGAGGAATTCGTGGGCTTATTGTTCGGTCATTG
GATTGGACCCTTCGCCGGTTTTCTGCGCATATAGTGCAGTGTAATTGGATTTACCAGAA
GTCTAGCGGACCCTCTTTACTATTCCCAGCTGTGATGGCGGTTTGTTGTGGTCCTACAA
GGGTCTTTGTGGACCGGGGTTTTTAGAATACGGACAATCCTTTGCCGATCGCCTGCGGC
GAGCGCCCCATCGGTTTGTGGTCAGAATATTGTCAATGCCATCGAGAGATCGGAGAATG
GATTGCGGATAAGGGTGGACTCGAGTTGGTCAAATTGCATTGGTACTCGACCAGTTCGT
GCACTATATGCAGAGCAATGATGAAGAGGATCAAGAT
```

(This sequence is written in Fasta format - see below for sequence formats.)

A computer could do it quicker, but it is still trivial to do it by eye. Especially as one of the sites has been picked out in bold. Can you find the other(s)? Sequence analyses impossible without a computer include, but are not limited to, most operations that involve the sequence databases. The DNA databases (Genbank EMBL DDBJ) are curated by three different groups in Bethesda, MD, Hinxton, UK and Mishima, JP but, because they exchange information on a daily basis, should be effectively the same in content. The DNA databases are doubling in size about every year; they currently (Oct 2008) comprise:

> 90 million sequences

and

99,116,431,942 base pairs

So finding all of the ecoRI sites in GenBank or even the whole of a printed copy of the human genome (3,200,000,000 bp) would take more than a few minutes.

This course will introduce you to some of the more commonly used bioinformatics tools; tell you how to use them and, more importantly, how to use them "correctly" or at least more effectively. Most of the analysis will be carried out on the World Wide Web

(WWW). This is partly because it is available to all comers without requiring direct access to the necessary computers, which serve as database and software repositories. But it is also partly because a well-designed Web site can be particularly user-friendly and intuitive in its operations.

There are likely to be network related problems trying to make 25 simultaneous connections over the Internet to the same site. Try doing the course exercises late in the evening, early in the morning (best for speed!) or at weekends.

This module in bioinformatics is designed to give you a flavour of what analytical and informative tools are available on the World Wide Web.

Software used in the course is many and varied. We have tried to put links to them all on the course website:

http://hpc.ilri.cgiar.org/training/BecA2012/Welcome.html

A few overall points for the course:

- Take the opportunity to compare and contrast different methods of doing a particular analysis.

- By all means take the defaults but be aware that changing them will almost certainly get more or better information.

- The Web is free, and you get what you pay for, so use the Web with care & caution.

- As with lab work it takes time to get the protocol working. Once you have one that works for you, write it down, bookmark and remember it. But note, the Web changes rapidly and you cannot afford to use outmoded technology for long.

- Where applicable we will also introduce you to the same tool implemented in the EMBOSS package. EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. EMBOSS integrates a range of currently available packages and tools for sequence analysis into a seamless whole. The EMBOSS package will be described in detail in a separate course module.

Introduction to Bioinformatics

Bioinformatics has been described as the storage, retrieval and analysis of biological sequence information. In this short course we will be taking a broader definition: how computers can maximise the biological information available to you. This will touch on determining the 3-D structure of bio-molecules and trying to relate this to their function as well as accessing the relevant literature. I hope that, by the end of the course, everyone will be adopting a more explicitly evolutionary understanding of 'their' molecule. The formal course practicals can be carried out entirely on the World Wide Web using Netscape or the other Web-browser. Nevertheless, we recommend using locally installed (FREE) software for the phylogenetic trees part of the course.

You should note that several important types of bioinformatic analysis are not freely accessible on the Web, but are available on various password controlled computers. In particular, types of analysis that require large amounts of computational power/time are best carried out off the web. Analyses of many genes are also often better done in an environment where a computer program does the pointing and clicking for you. For the record, the EMBOSS package is a suite of programs which carry out almost all the analyses that a molecular biologist might want to do with/on DNA or protein sequences (secondary structure prediction, two sequence alignment, conceptual translation of DNA, restriction site analysis, primer design, as well as homology searching, multiple sequence alignment etc.). For phylogenetic inference and tree drawing, the PHYLIP package (versions available for PCs, Macs and Unix) will answer most needs. Both of these software packages and a variety of other sequence analysis packages are available for download from the Internet.

The web, by contrast, is a total mess: the same program is implemented with different defaults at different sites; it is often not clear what those defaults, options and parameters are; the results are not easily transferred to a different program. So it is free, but there is a cost! You are advised to validate any analysis against the results yielded by other sites.

For a good introduction to Bioinformatics, read the first chapter of Developing Bioinformatics Computer Skills, Cynthia Gibas & Per Jambeck, available online at http://oreilly.com/catalog/bioskills/chapter/ch01.html

Databases

Databases are of course the core resource for bioinformatics. There is plenty of software for analysing one or a few sequences, but many of the computationally interesting and biologically informative programs access databases of information. Frequently used classes are the biological sequence databases. These include:

- EMBL (European Mol Biol Lab)
- GenBank
- DDBJ (DNA DB of Japan)

These three DNA databases exchange their data on a daily basis and so should be identical as to content. They are, however, rather different in format:
Each of the database cited above consists of a (very large number) of entries, each consisting of a single sequence preceded by a quantity of 'annotation' that puts the sequence in its biological, functional and historical context. Without the annotation, GenBank would be a meaningless string of 32 billion As Ts Cs and Gs. Compare and contrast the two extracts from a) EMBL and b) Genbank (DDBJ has the same look-and-feel as Genbank):

a) EMBL

```
ID ECRECA standard; DNA; PRO; 1391 BP.
AC V00328; J01672;
DT 09-JUN-1982 (Rel. 01, Created)
DT 12-SEP-1993 (Rel. 36, Last updated, Version 4)
DE E. coli recA gene.
KW .
OS Escherichia coli
OC Bacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
OC Escherichia.
RN [1]
RP 1-1374
RX MEDLINE; 80234673.
RA Sancar A., Stachelek C., Konigsberg W., Rupp W.D.;
RT "Sequences of the recA gene and protein";
RL Proc. Natl. Acad. Sci. U.S.A. 77:2611-2615(1980).
```

b) GenBank

```
LOCUS ECRECA 1391 bp DNA BCT 12-SEP-1993
DEFINITION E. coli recA gene.
ACCESSION V00328 J01672
KEYWORDS .
SOURCE Escherichia coli.
ORGANISM Escherichia coli
Eubacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
Escherichia.
REFERENCE 1 (bases 1 to 1374)
AUTHORS Sancar,A., Stachelek,C., Konigsberg,W. and Rupp,W.D.
TITLE Sequences of the recA gene and protein
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 77 (5), 2611-2615 (1980)
```

You can see that these two are obviously talking about the same sequence from *E.coli*, but the information is encoded in a rather different way. This makes no difference to us reading the text, but causes problems when writing a program to interrogate a database.

Each database entry has a name, called ID or LOCUS, which tries to be mnemonic and marginally informative. More importantly each has an accession number which is arbitrary but which remains attached to the sequence for the rest of time. The organism might become reclassified, the gene may get renamed and the ID is thus subject to change, but by noting the accession number you should always be able to identify and retrieve the sequence. Note also that the original publication is cited. Usually there will be other papers documenting functional analysis, mutations, allelic variations, 3-D structure and so on.

Further down in the entry is annotation about the sequence itself, so that the sequence is parsed into meaningful bits called a features table:

a) EMBL

```
FT source 1. .1391
FT /organism="Escherichia coli"
FT /db_xref="taxon:562"
FT mRNA 191. .>1391
FT /note="messenger RNA"
FT RBS 229. .233
FT /note="ribosomal binding site"
FT CDS 239. .1300
FT /db_xref="SWISS-PROT:P03017"
FT /transl_table=11
FT /gene="recA"
FT /product="recA gene product"
FT /protein_id="CAA23618.1"
FT
FT mutation 353. .353
FT /note="g to a in recA441 (E to K)"
FT mutation 720. .720
FT /note="g to a in recA1 (G to D)"
```

b) GenBank

```
FEATURES Location/Qualifiers
source 1..1391
/organism="Escherichia coli"
/db_xref="taxon:562"
mRNA 191..>1391
/note="messenger RNA"
RBS 229..233
/note="ribosomal binding site"
gene 239..1300
/gene="recA"
CDS 239..1300
/gene="recA"
/codon_start=1
/transl_table=11
/product="recA gene product"
/db_xref="SWISS-PROT:P03017"
mutation 353
/gene="recA"
/note="g to a in recA441 (E to K)"
mutation 720
/gene="recA"
/note="g to a in recA1 (G to D)"
```

Again you can see that the information exchange between Genbank and EMBL includes all significant portions of the annotation. Such useful signals and data as the open reading frame (CDS for CoDing Sequence), the ribosome binding site, intron boundaries, signal peptides, variants/mutations may be recorded.

Protein databases:

- SwissProt
- PIR (Protein Information Resource)

- GenPept

a) Swissprot

```
ID RECA_ECOLI STANDARD; PRT; 352 AA.
AC P03017; P26347; P78213;
DT 21-JUL-1986 (REL. 01, CREATED)
DT 21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT 15-DEC-1998 (REL. 37, LAST ANNOTATION UPDATE)
DE RECA PROTEIN.
GN RECA OR LEXB OR UMUB OR RECH OR RNMB OR TIF OR ZAB.
OS ESCHERICHIA COLI, AND SHIGELLA FLEXNERI.
OC BACTERIA; PROTEOBACTERIA; GAMMA SUBDIVISION; ENTEROBACTERIACEAE;
OC ESCHERICHIA.
...
...
CC -!- FUNCTION: RECA PROTEIN CAN CATALYZE THE HYDROLYSIS OF ATP IN THE
CC PRESENCE OF SINGLE-STRANDED DNA, THE ATP-DEPENDENT UPTAKE OF
CC SINGLE-STRANDED DNA BY DUPLEX DNA, AND THE ATP-DEPENDENT
CC HYBRIDIZATION OF HOMOLOGOUS SINGLE-STRANDED DNAS. IT INTERACTS
CC WITH LEXA CAUSING ITS ACTIVATION AND LEADING TO ITS AUTOCATALYTIC
CC CLEAVAGE.
CC -!- INDUCTION: IN RESPONSE TO LOW TEMPERATURE. SENSITIVE TO
CC TEMPERATURE THROUGH CHANGES IN THE LINKING NUMBER OF THE DNA.
CC -!- DATABASE: NAME=E.coli recA Web page;
CC WWW="http://monera.ncl.ac.uk:80/protein/final/reca.htm".
KW DNA DAMAGE; DNA RECOMBINATION; SOS RESPONSE; ATP-BINDING; DNA-
BINDING;
KW 3D-STRUCTURE.
FT INIT_MET 0 0
FT NP_BIND 66 73 ATP.
FT CONFLICT 112 112 D -> E (IN REF. 5).
FT TURN 4 4
FT HELIX 5 21
FT HELIX 23 25
FT TURN 29 30
```

etc
etc

b) PIR

```
>P1;RQECA
recA protein - Escherichia coli
C;Species: Escherichia coli
C;Date: 31-Jul-1980 #sequence_revision 14-Nov-1997 #text_change 14-Nov-1997
C;Accession: G65049; A93847; A93846; S11931; S63525; S63979; A03548
...
C;Comment: The recA protein plays an essential role in homologous recombination,
in induction of the SOS response, and in initiation of stable DNA replication.
C;Genetics:
A;Gene: recA
A;Map position: 58 min
C;Superfamily: recA protein
C;Keywords: ATP; DNA binding; DNA recombination; DNA repair; P-loop; SOS
response
F;67-75/Region: nucleotide-binding motif A (P-loop)
F;141-145/Region: nucleotide-binding motif B
F;73/Binding site: ATP (Lys) #status predicted
```

Note that these two entries refer to the same gene from E.coli despite differences in the way the data is encoded. However, in contrast to the difference between EMBL and Genbank, the quality of the annotation is quite different. The 3-D structure of this gene has been worked out and this information is reflected in the SwissProt entry as the position of every alpha-helix and beta-sheet is noted. In general, the quality of the annotation and the minimization of internal redundancy makes SwissProt the preferred database to use. However, note that PIR records the Genetic Map position of the gene; so it is probably good to scrutinize both databases to abstract maximal information.
SwissProt also gives added value by incorporating a large number of DR (database reference) tags, pointing to equivalent information in other databases.

a) SwissProt:

```
DR EMBL; V00328; G42673; -.
DR EMBL; X55553; -; NOT_ANNOTATED_CDS.
DR EMBL; AE000354; G1789051; -.
DR EMBL; D90892; G1800085; -.
DR PIR; A03548; RQECA.
DR PIR; S11931; S11931.
DR PDB; 1REA; 31-OCT-93.
DR PDB; 2REB; 31-OCT-93.
DR PDB; 2REC; 01-APR-97.
DR PDB; 1AA3; 23-JUL-97.
DR SWISS-2DPAGE; P03017; COLI.
DR ECO2DBASE; C039.3; 6TH EDITION.
DR ECOGENE; EG10823; RECA.
DR PROSITE; PS00321; RECA; 1.
DR PFAM; PF00154; recA; 1.
```

When these are used as hypertext links they can enable a WWW browser to locate an extraordinary depth of detail about a given entry, 3-D structure (PDB), protein motifs (Prosite), families of related genes (Pfam), the DNA sequence (EMBL) and a couple of specialist E.coli added-value databases. SRS is one program that makes these hypertext links. The PIR cross-references are far fewer and less explicit; its reference to Genbank (GB:U00096) refers to the whole E.coli genome, whereas SwissProt points specifically to the gene (DR EMBL; V00328)

b) PIR

```
...
A;Cross-references: GB:AE000354; GB:U00096; NID:g2367149; PID:g1789051;
UWGP:b2699
```

All these databases are made up of entries, concatenated one after the other in plain readable text. As such they are far bigger than necessary if you are trying to analyze the sequence rather than interrogate or browse the annotation. For these purposes, special high-compressed databases can be constructed. Frequently these are not readable by humans because they have been optimized for speed reading computers.
One of the simplest compression protocols is called Fasta format in which the annotation is edited down to a single title line followed by the sequence. The sequence at the top of the chapter is in Fasta format. All protein databases use the one-letter amino acid code, can you think why this might be?

Sequence Related Databases

Not all biologically relevant Databases consist of sequences and annotation. There are databases of journal abstracts, taxonomy, 3-D structures, mutations and metabolic pathways. Some of the most useful of these are databases which specialise in particular entities that can be found dispersed in the "whole sequence" databases.

You notice one of the cross-references for the SwissProt entry is:

DR PROSITE; PS00321; RECA; 1.

Prosite is a database of protein motifs. PS00321 is a family of proteins that all have the motif:
PA A-L-K-F-[FY]-[STA]-[STAD]-[VM]-R

and are all believed to bind DNA, hydrolyze ATP and act as a recombinase. One of the members of this family is the recA gene in E.coli which gives its name to PS00321. In the pattern above, the residues within [square brackets] are alternatives. Convince yourself that ALKFFAAVR could belong to the family but ALKFAAAVR could not. There are more than 1000 other families classified in a similar way. Finding a Prosite link in a SwissProt gene is a great help in finding other proteins related by structure and/or function.

Interpro - http://www.ebi.ac.uk/interpro/

You should also be aware of the Interpro project which incorporates and sorts data from a diversity of protein motif and domain databases into one searchable meta-database.

Sequence formats

As we have seen comparing database entries above, there are dozens of different ways in which you can store or represent the same fundamental information. Databases are often compiled in, highly conventionalized, readable English text. Computers, being not so bright, will have difficulty reading and interpreting the information unless the conventions are quite rigidly obeyed. There are a very large number of ways you can write, store and transmit simple one-dimensional sequence files. A common sequence interchange program called 'readseq' recognizes at least 22 different file formats. If a computer program does not recognize the format of an input sequence it may not work or, worse, misinterpret header lines as sequence data or otherwise mangle your analysis. The EMBOSS package can also convert between different sequence formats.

EMBOSS

seqret
Reads and writes (returns) sequences in different formats. It can also read in a sequence from a database and write it to a file.

Some commonly used file/sequence formats are shown below:

1) Fasta (named for a widely used homology searching program) – single title line beginning >:

```
>ECRGCG TRANSLATE of: ecrgcg 1 to: 1062
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

2) Staden (named after Rodger Staden - early, but still extant, software writer) – same as raw sequence:

```
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

3) NBRF/PIR (named after the protein database):

```
>P1;ecrgcg.pep
ecrgcg.pep, 354 bases, 218 checksum.
MAIDENKQKA LAAALGQIEK
ALGAGGLPMG RIVEIYGPES
TPKAEIEGE*
```

Accession numbers

The information above makes you aware of the diversity of ways in which something so simple as a one-dimensional sequence may be represented. Another source of confusion is the variety of identifying numbers attached to sequences and knowing to which database they refer. Accession numbers are used as unique and unchanging numbers. They are not mnemonic, although databases also have a less stable, more memorable nomenclature: HBB_HUMAN, HSHBB, HUMHBB 2HBB are all human beta globin IDs in various databases,

- **GenBank/EMBL** accession numbers: originally a letter followed by 5 digits (X32152, M22239). When the number of sequences exceeded 2,600,000 - 2 letters followed by 6 digits (AL234556, BF345788).
- **SwissProt**. Still one letter followed by 5 digits, letter is either O,P,Q. P23445.
- **PIR**: the 'other' protein database, one letter followed by 5 digits, but numbers confusable with EMBL/GenBank: B93303 is chimp haemoglobin in PIR but a random genomic clone fragment in EMBL.
- **GenPept**. Conceptual translations from DNA that have not yet been annotated well enough to get into SwissProt. three letters and five digits, e.g.: AAA12345.
- **Trembl** (Translated EMBL): O, P or Q followed by 5 letters/digits.
- **PDB** protein structure records: 1 digit and three letters 1HBA, 1TUP

More recently, an attempt has been made to reduce the redundancy in the databases (there were 180 copies of D. melanogaster alcohol dehydrogenase each with its own accession number). One result is RefSeq - NCBI's "reference sequence" database

RefSeq: Two letters, and underscore bar, and six digits,
mRNA records (NM_*) NM_000492 **genomic DNA contigs** (NT_*) NT_000347 **curated/annotated Genomic regions** (NG_*) NG_000567 **Protein sequence records** (NP_*) NP_000483

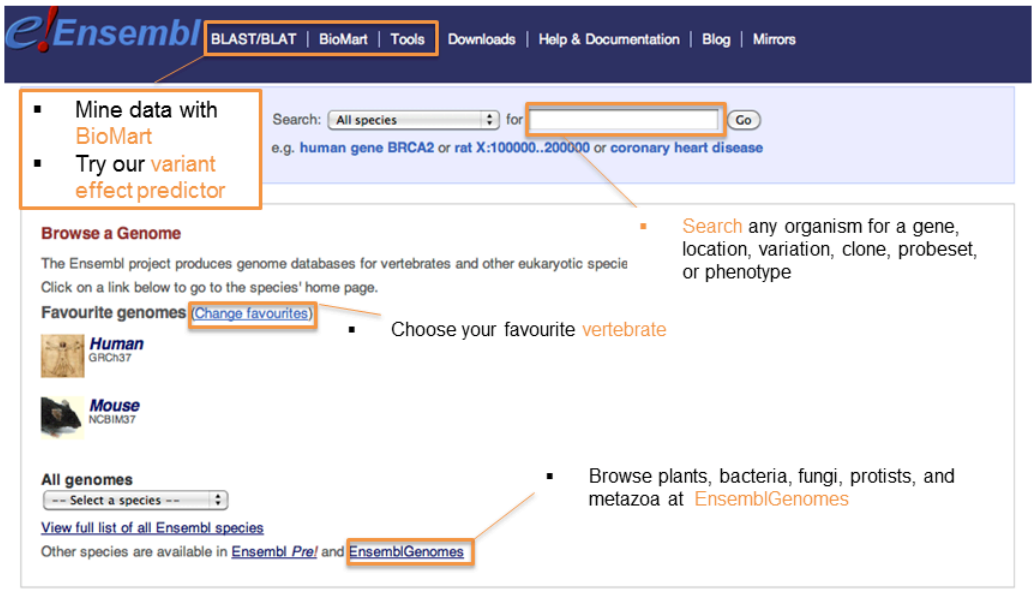
We will see how RefSeq is becoming the central resource for gene characterization, expression studies, and polymorphism discovery. Because of the high level of necessary curation, it is not anywhere close to being comprehensive even for those species that are included.

Accession numbers give the community a unique label to attach to a biological entity, so we all know we are talking about the same thing. Sequences in databases evolve as their real biological counterparts do. They need to be updated, corrected and merged and we need to know which version of the sequence entry is being referred to. GenBank has used gi numbers and, more recently, version numbers for this. Each small change made to a Genbank record gets the next gi number e.g. gi6995995 and so is totally arbitrary. Version numbers are appended to the accession number after a dot – V00234.2, NM_000492.2.

Interrogating (sequence) databases

Ensembl – http://www.ensembl.org

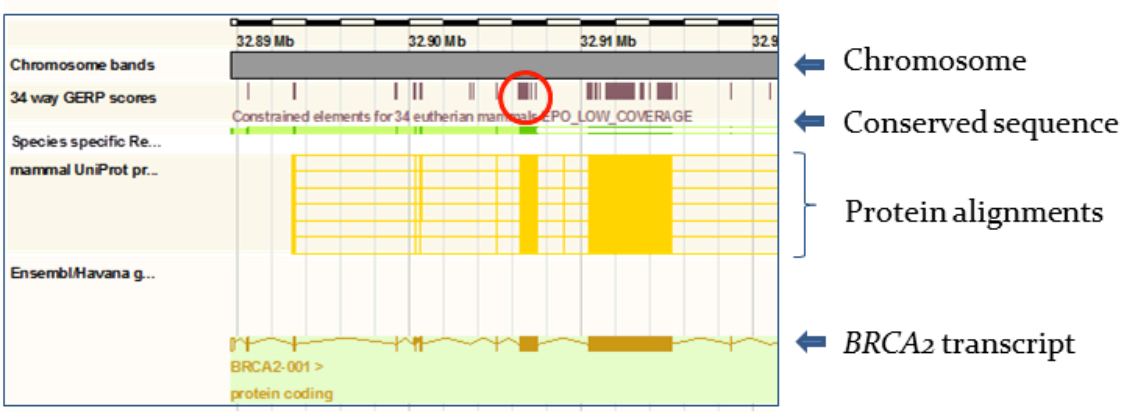
Ensembl provides genes and other annotation such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in UniProtKB and NCBI RefSeq databases, along with manual annotation from the VEGA/Havana group. All the data are freely available and can accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (Perl API). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the BioMart web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user!


e!Ensembl
BLAST/BLAT | BioMart | Tools
Downloads | Help & Documentation | Blog | Mirrors
Mine data with BioMart
Try our variant effect predictor
Search: All species for Go
e.g. human gene BRCA2 or rat X:100000..200000 or coronary heart disease
Search any organism for a gene, location, variation, clone, probeset, or phenotype
Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic specie
Click on a link below to go to the species' home page.
Favourite genomes (Change favourites)
Choose your favourite vertebrate
Human GRCh37
Mouse NCBIM37
All genomes
-- Select a species --
View full list of all Ensembl species
Other species are available in Ensembl Pre! and EnsemblGenomes
Browse plants, bacteria, fungi, protists, and metazoa at EnsemblGenomes

Introduction to Ensembl

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates chordate genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'compara' team. Most annotation is updated every two months, leading to

increasing Ensembl versions (such as version 62), however the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non---chordates, namely bacteria, plants, fungi, metazoa, and protists.

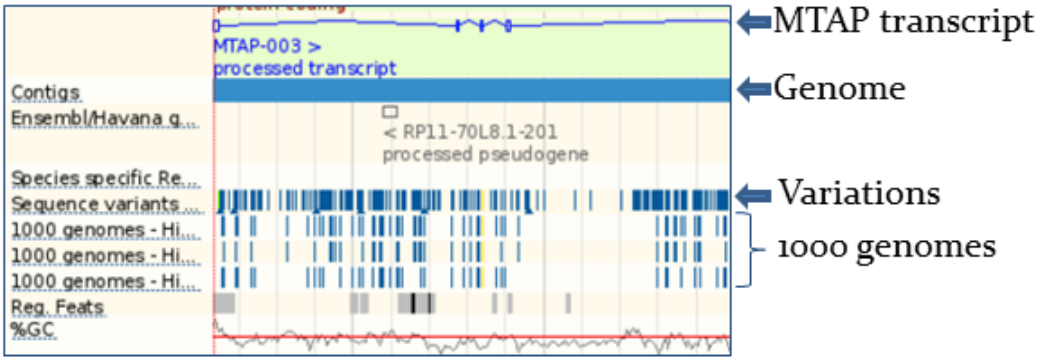


The region in detail view

The vast amount of information associated with the genomic sequence demands a way to organise and access that information. This is where genome browsers come in. Ensembl strives to display many layers of genome annotation into a simplified view for the ease of the user. The picture above shows the 'Region in Detail' page for the BRCA2 gene in human. The example shows blocks of conserved sequence reflecting conservation scores of sequence identity on a base pair level across 34 species. Conserved regions are displayed as dark blocks that represent local regions of alignment. One of the blocks is circled in red. You would only have to click on this block to see more details.

Also in this figure are proteins from the UniProtKB aligned to the same genomic region. Filled yellow blocks show where these UniProtKB proteins align to the genome, and gaps in the alignment are shown as empty yellow blocks. Note, in this case, the UniProtKB proteins support most of the exons shown in the Ensembl BRCA2---001 transcript (in gold).

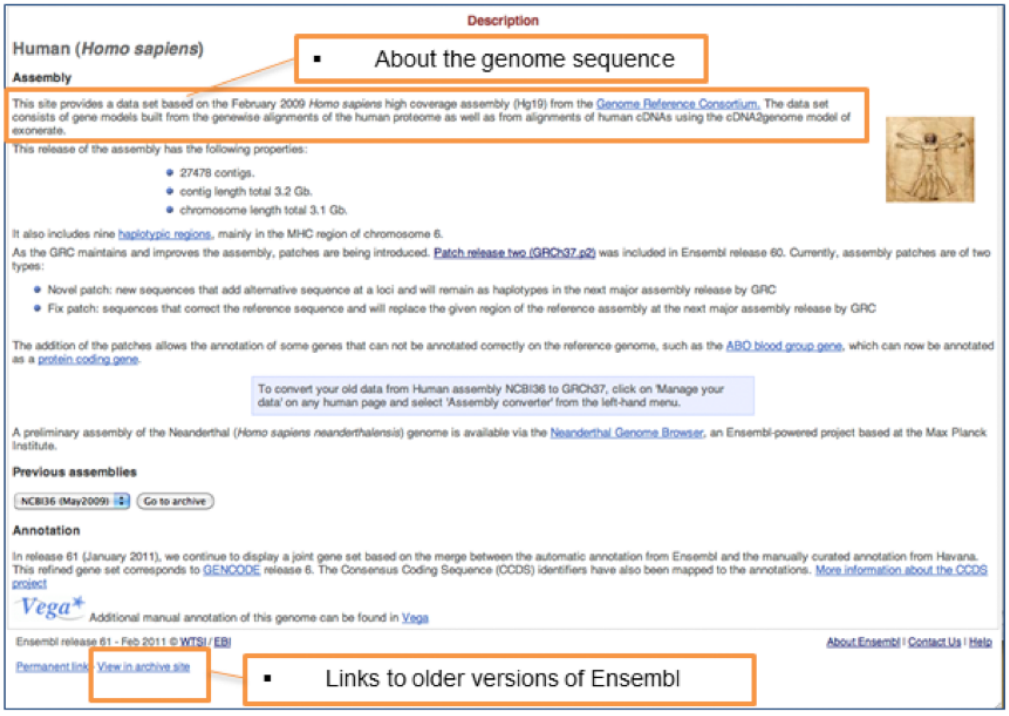
Both Ensembl and Vega (Havana) transcripts are portrayed as exons (boxes) and introns (connecting lines). In fact, filled boxes show coding sequence, and empty boxes reflect UnTranslated Regions (UTRs). This 'Region in Detail' view is useful for comparing Ensembl gene models with current proteins and mRNAs in other databases like NCBI RefSeq, EMBL---Bank, and, in the example above, UniProtKB. Everything in this view is aligned to the genome.

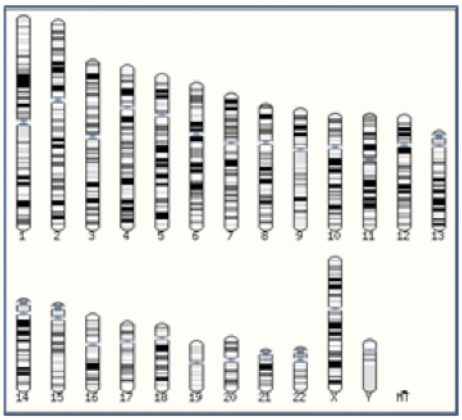


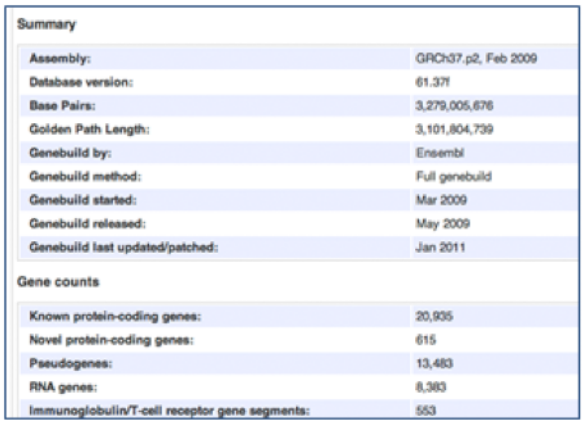
The region in detail view: 1000 genomes track

The region in detail view can be configured (using the Configure this page tool button) to show regulatory features, sequence variation, and more! Click on any vertical line in the variation track for a menu about the SNP (single nucleotide polymorphism) or InDel (insertion deletion mutation). Clicking on 'Variation properties' in the pop---up box will bring you to an information page for the genetic variation, including links to population frequencies, if known. You can do the same for any regulatory feature.

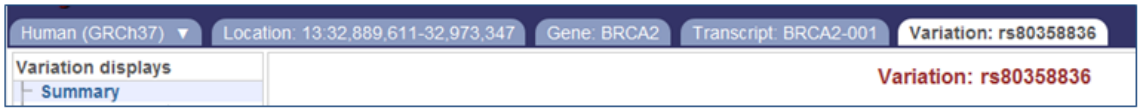
An index page is provided for each species with information about the source of the genomic sequence assembly, a karyotype (if available), and a link to past or archive sites. The picture below shows the Ensembl homepage for human. Links to the human karyotype, a summary of gene and genome information, and the most common InterPro domains in the genome are found at the left of this index page.


Description
Human (Homo sapiens)
Assembly
About the genome sequence
This site provides a data set based on the February 2009 Homo sapiens high coverage assembly (Hg19) from the Genome Reference Consortium. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.
This release of the assembly has the following properties:
27478 contigs.
contig length total 3.2 Gb.
chromosome length total 3.1 Gb.
It also includes nine haplotypic regions, mainly in the MHC region of chromosome 6.
As the GRC maintains and improves the assembly, patches are being introduced. Patch release two (GRCh37.p2) was included in Ensembl release 60. Currently, assembly patches are of two types:
Novel patch: new sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC
Fix patch: sequences that correct the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC
The addition of the patches allows the annotation of some genes that can not be annotated correctly on the reference genome, such as the ABO blood group gene, which can now be annotated as a protein coding gene.
To convert your old data from Human assembly NCBI36 to GRCh37, click on 'Manage your data' on any human page and select 'Assembly converter' from the left-hand menu.
A preliminary assembly of the Neanderthal (Homo sapiens neanderthalensis) genome is available via the Neanderthal Genome Browser, an Ensembl-powered project based at the Max Planck Institute.
Previous assemblies
NCBI36 (May2009)
Go to archive
Annotation
In release 61 (January 2011), we continue to display a joint gene set based on the merge between the automatic annotation from Ensembl and the manually curated annotation from Havana. This refined gene set corresponds to GENCODE release 6. The Consensus Coding Sequence (CCDS) identifiers have also been mapped to the annotations. More information about the CCDS project
Vega
Additional manual annotation of this genome can be found in Vega
Ensembl release 61 - Feb 2011 © WTSI / EBI
About Ensembl | Contact Us | Help
Permanent link
View in archive site
Links to older versions of Ensembl





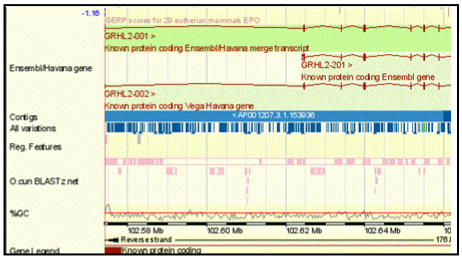
Ensembl devotes separate pages and views in the browser to display a variety of information types, using a tabbed structure.



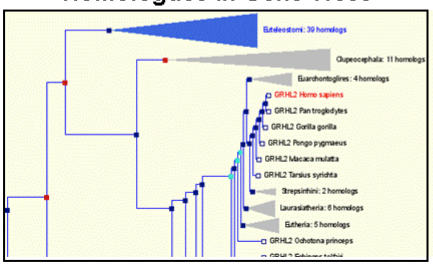
View genotype information in the variation tab, gene trees in the gene tab, a chromosomal region in the location tab, and cDNA sequence alongside the protein translation in the transcript pages. Compare conserved regions with the position of genes and population variation in the Region in Detail view. See homology relationships in the gene page, or perform a BLAST or BLAT search against any species in Ensembl.

Transcript Sequence w/Variations

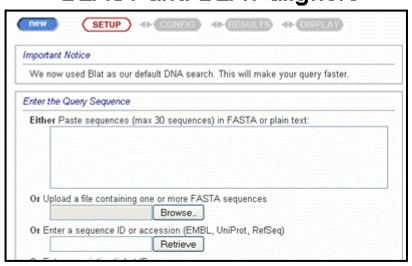
Genes, SNPs, and Conserved Regions



Homologues in Gene Trees



BLAST and BLAT aligners


new
SETUP
CONFIG
RESULTS
DISPLAY
Important Notice
We now used Blat as our default DNA search. This will make your query faster.
Enter the Query Sequence
Either Paste sequences (max 30 sequences) in FASTA or plain text:
Or Upload a file containing one or more FASTA sequences
Browse...
Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)
Retrieve

Ensembl Exercises

1. Open the home page of Ensembl. Search for the human gene GFAP. Select the ensembl gene ENSG00000131095.

2. How many transcripts does this gene have?

3. What is the genomic location of the gene?

4. What is the length and how many exons do the different transcripts have?

5. Choose the first transcript that has supporting evidence. Copy the sequence for the first exon including the UTR (Un-translated region). Submit the sequence as an answer to this question!

6. Now use Blat and search with this same sequence of that first exon against the human genome. On the result click on [C] (ContigView) for the alignment that is 100% identical. Does your Blat hit correspond to the first exon? Export the picture and submit it as an answer to this question!

7. Again choose the same transcript that has the supporting evidence. Please find the protein translation for this transcript. Copy and submit it as an answer to this question! What is the protein length? (How many residues?)

8. Click on the gene tab. How many paralogues and how many orthologues has been found for this gene? Explain the difference between orthologues and paralogues genes!

9. Click on Orthologues! Can you find a dog orthologue for this gene? If so what is the chromosome location in the dog genome?

10. Click on the transcript tab and then click on Protein ID. . What can you find about the protein family of this gene? Click on 'Domains & Features' right arrow. Follow other links that might interest you and find out more about this gene!

SRS - http://srs.ebi.ac.uk/

The DNA databases are enormously rich information resources partly because they are so big, but it would make little sense if it consisted of a long list of As Ts Cs and Gs. There are millions of individual entries in EMBL. An entry could be a fragment as short as 3 base pairs (e.g. M23994) or a large contig consisting of many genes, including complete eukaryotic chromosomes (e.g. X59720). The value of the database lies substantially in the quality of the annotation that puts the sequence in its biological context.

As a biologist you may need to be able to interrogate the Database to find particular sequences or a set of sequences matching given criteria, such as:

The sequence published in Cell 31: 375-382
All sequences from Aspergillus nidulans
Sequences submitted by Peter Arctander
Flagellin or fibrinogen sequences
The glutamine synthase gene from Haemophilus influenzae
The upstream control region of Bacillus subtilis Spo0A

SRS (Sequence Retrieval System) is a very powerful, WWW-based tool, developed by Thure Etzold at EMBL and subsequently managed by Lion Biosciences, for interrogating databases and abstracting information from them.

One of the neatest features of SRS is the fact that interrelated databases can be cross-referenced with WWW hypertext links. This means that you can discover the protein sequence, the cognate DNA sequence, a family of related proteins in other species, a Medline reference to read an abstract of the original publication, a 3-D structure - all with a few point-and-clicks with the mouse.

There are several SRS servers on the Web. We will be using

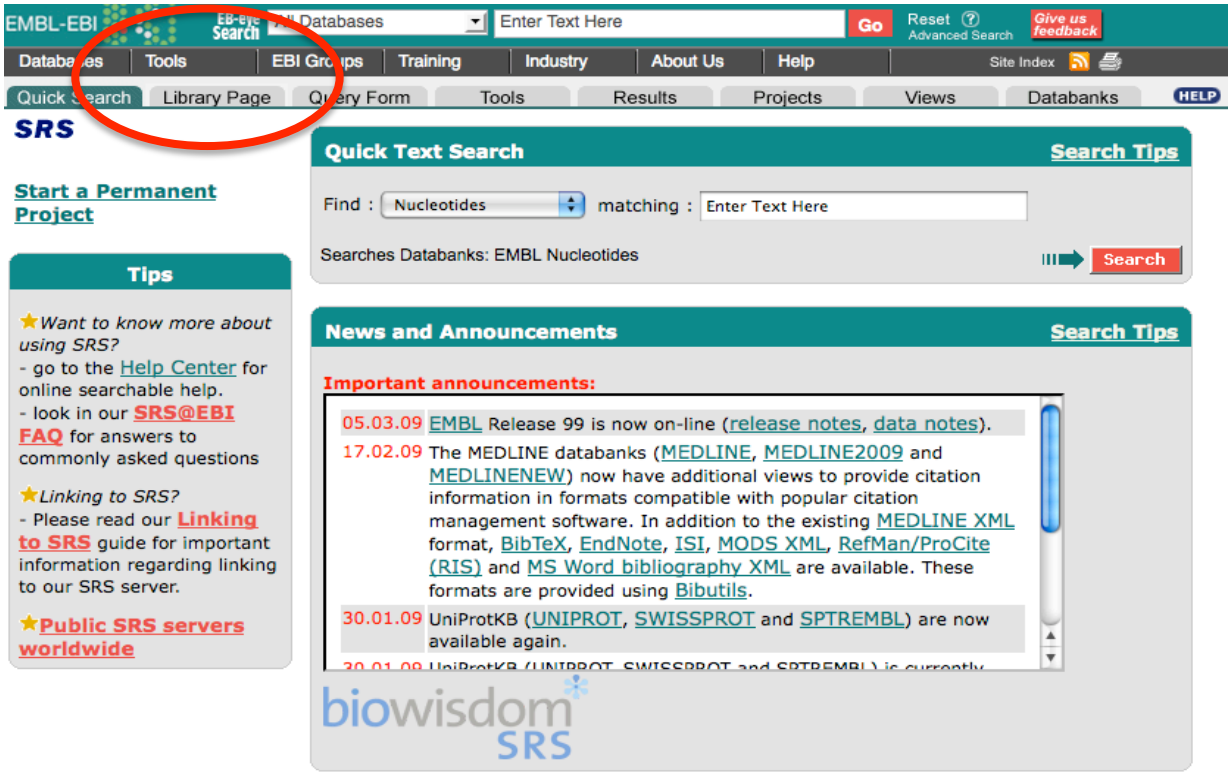
http://srs.ebi.ac.uk/

at the EBI in England because a) it has a large number of interlinked databases b) connectivity to the UK is good c) they are attempting to interconnect their SRS server with their clustalW server and blast server.

With experience and practice you will get to use as much of SRS's power as necessary to obtain the results you need. Below, as a worked example, a series of instructions to obtain the sequences of serum resistance associated proteins in *Trypanosoma brucei* in SwissProt, and download them locally to carry out a multiple sequence alignment using, say, clustalW. It should also be possible to do the multiple alignments on the EBI clustalW server.

Use your browser (Netscape?) to go to http://srs.ebi.ac.uk/ or one of the other SRS servers at the top of the Course page. You should see the following options:

Click on '**Library Page'**.


EMBL-EBI
Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Tools
Training
Industry
About Us
Help
Site Index
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
HELP
SRS
Start a Permanent Project
Quick Text Search
Search Tips
Find :
Nucleotides
matching :
Enter Text Here
Searches Databanks: EMBL Nucleotides
Search
Tips
Want to know more about using SRS?
- go to the Help Center for online searchable help.
- look in our SRS@EBI FAQ for answers to commonly asked questions
Linking to SRS?
- Please read our Linking to SRS guide for important information regarding linking to our SRS server.
Public SRS servers worldwide
News and Announcements
Search Tips
Important announcements:
05.03.09 EMBL Release 99 is now on-line (release notes, data notes).
17.02.09 The MEDLINE databanks (MEDLINE, MEDLINE2009 and MEDLINENEW) now have additional views to provide citation information in formats compatible with popular citation management software. In addition to the existing MEDLINE XML format, BibTeX, EndNote, ISI, MODS XML, RefMan/ProCite (RIS) and MS Word bibliography XML are available. These formats are provided using Bibutils.
30.01.09 UniProtKB (UNIPROT, SWISSPROT and SPTREMBL) are now available again.
biowisdom SRS

database(s) that you wish to search. The databases may be of various types, including:

Sequence: Swissprot, sptrembl, PIR (Protein) or EMBL, emblnew (DNA)

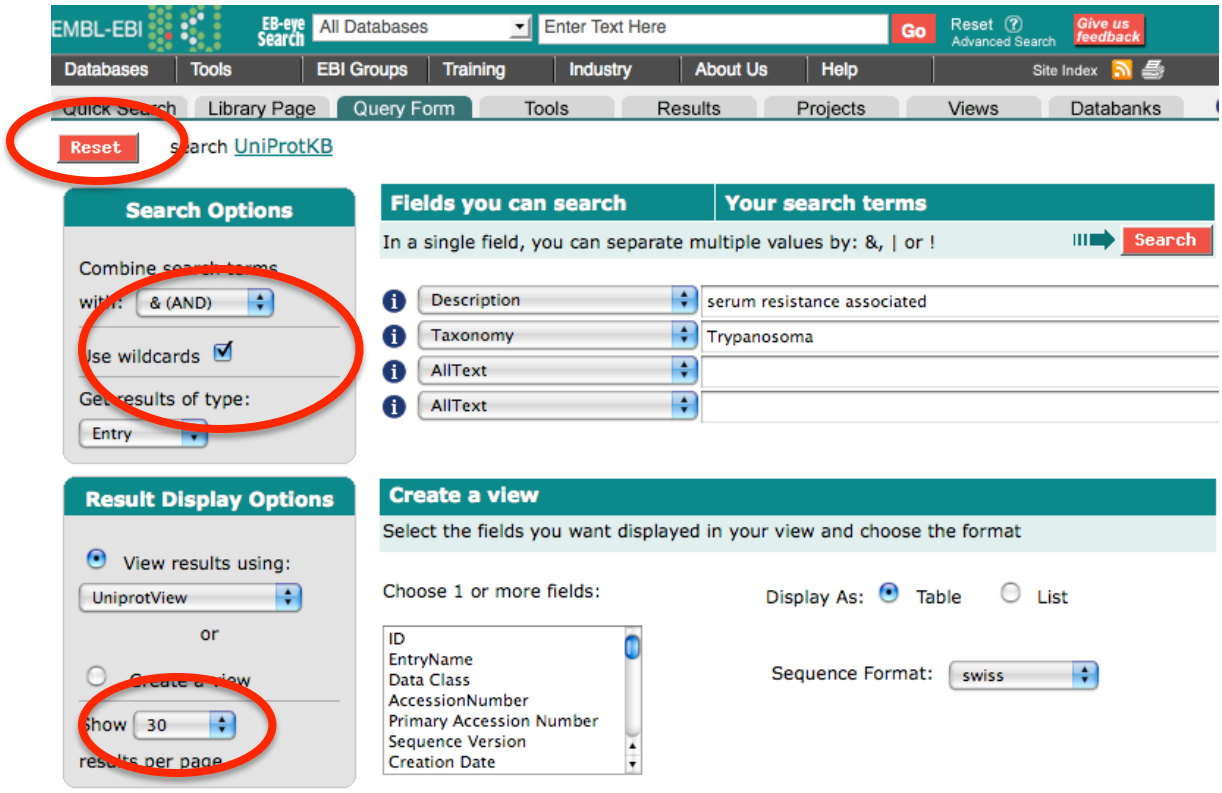
Sequence related: prosite, blocks, prints (protein motifs and alignments), repbase (restriction enzymes),

Protein3Dstructure: PDB, HSSP

For more information about the contents of the database click on the relevant blue underlined hypertext link - UniProt say.

- **Click the box [_] to the left of UniProtKB.**
- **Click on the Query Form tab at the top of the page**

This will move you to a Query Form Page that permits you to submit particular queries (such as have been suggested at the beginning of this chapter) to the databases. At the top of this page will be a note of which database(s) you have chosen to search and a block of four text-insert boxes which you can use to enter your question.



to the left you will see some things you can change including:

1. **[Reset]** - which clears the screen.
2. **combine search terms &(AND)** - which enables you to apply other logical (boolean) operators.
3. **Use wildcards** which means that "bact" will be interpreted as bact* and look for bacteria, bacteriophage, etc.
4. **Number of entries to display per page** (default is 30)

Your question can be entered into one of more of the text-insert boxes, thus:

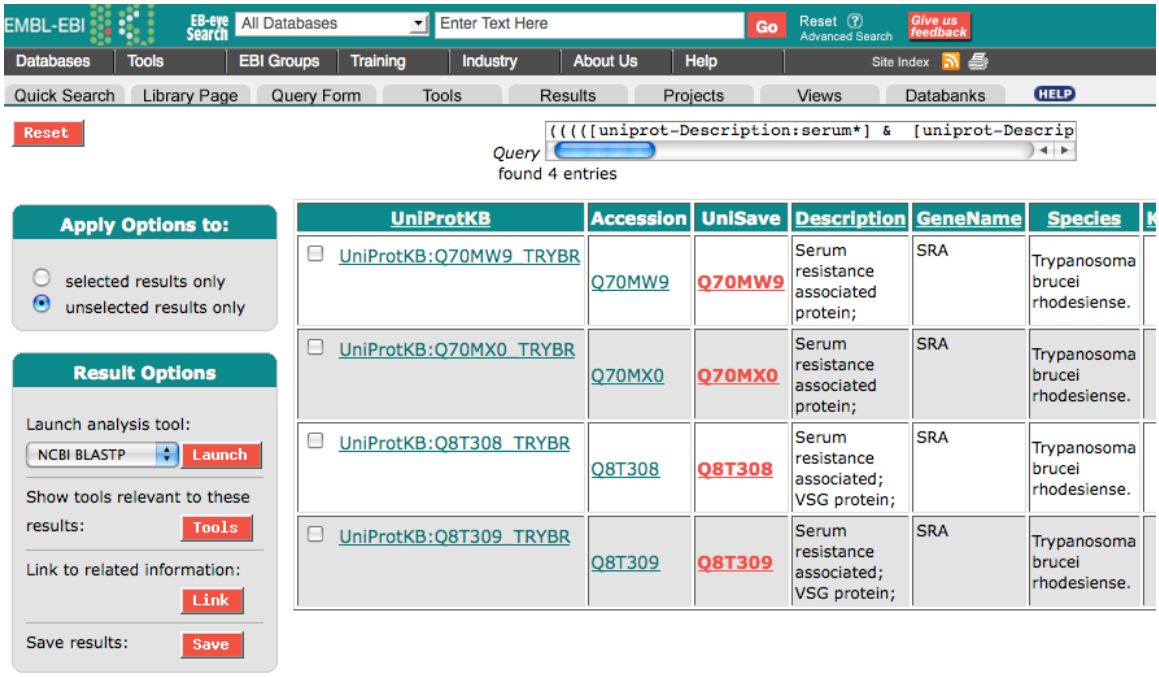
- **Click [All text] change to [Description]** and **insert "serum resistance associated" in box**

Note: it does not have to be "serum resistance associated" it could be ubiquitin or haemoglobin or hemoglobin or actin & alpha. Separate keywords in the same box have to be linked by a logical (Boolean) operator such as

and: &
or: |
but not: !

- **Click the next [All text] change to [Taxonomy] and insert "Trypanosoma" in box**
- **Click Search**

a new window appears with Query "([uniprot-Description: serum resistance associated*] & [uniprot-Taxonomy:Trypanosoma*] " found 4 entries. This is how SRS interprets what you have entered in the boxes and the numbers of "hits" found.



- **Under Display options change [UniprotView] to [FastaSeqs]**
- **Click [Save]**
- **Save as type – Text File (*.txt).**
- **Click [Save]**
- **Change selection .../wgetz to .../serum.pro and then Click Save.**

This should dump the concatenated fasta format protein sequences into a local file called serum.pro. You can use this file as input for clustalW. There may be local security difficulties with downloading sequences onto a public terminal - check with your neighbours or your demonstrator.

Query manager: a powerful tool

A quick example will show how you can combine very complex queries to zero in on the sequence(s) you need.
Having selected your database(s) go to the Query Form Page and enter:

- **[Description] calmodulin**

you should get about 1140 entries.

- **Click [QUERY] tab** at the top of the page to get a new page and enter:
- **[Organism name] human** (or indeed Homo sapiens)

this will get you a large number of sequences.

- **Click [RESULTS] tab** at the top of the page

A new window should appear with the results for all the queries you have entered in the current SRS session. In the top box of this page **enter "Q1 & Q2"** (leave off the quotes!) **Note**: Your mileage may vary here. Q1 and Q2 may refer to earlier queries in this SRS session (osteonectin?) so use good judgement.
You have just used a boolean logical expression to yield sequences which are a) human and b) have "calmodulin" in the SwissProt description. This shows you how it can be unreliable to depend on the annotation to get homologous sequences. Nevertheless, the list should contain the SwissProt entry for CALM_HUMAN.

Questions

1. Can you think of a better way to find other mammalian calmodulin genes?

2. If you do a search in SwissProt for "calmodulin" using the [AllText] descriptor instead of [Description] you find many more entries, why do you think you get more entries under this search?

3. There are more entries in SwissProt under [Organism] dog than [Author] dog, but more for [Author] wolf than [Organism] wolf. Why do you think this is so?

4. Searching [Organism] mouse in SwissProt yields some plant sequences: prove this by finding sequences matching [Organism] mouse & [Taxon] viridiplantae. Why is this so? (Clue: append wildcard *).

You should be able to reveal the full SwissProt entry for any protein sequence. If you do this you will see several (? blue, underlined) hypertext links to related databases. Almost certainly at least one of these will be EMBL and one to Medline. Probably one will be the prosite motif database. If the 3-D structure is known, one link will be to PDB. Investigate these other databases to get as much relevant information as possible about your sequence.

Aside: Displaying 3-D structures is not “fitted as standard” on all terminals. You may need to get a copy of the RasMol 3-D structure viewer and install it in such a way that your Netscape/IE will recognise it and connect suitable (3-D sequence) file to it.

To display a PDB entry of 3-D coordinates as a rotatable, colorable model you need to click on the **[save]** button. The change the "use mime type" choice-box to **chemical/x-pdb** and then click on the **[save]** box. This should fire up CHIME a WWW implementation of RasMol.) Your mileage may vary!

It is this, interlinked databases, aspect of SRS which gives it a large part of its power. You can extend your search to include other sequences related in some particular (or peculiar!) way. The Prosite link allows you to find members of a protein family. The EMBL link allows you to find the introns and the intron splice junctions, not to mention the ribosome-binding site, the stop codon and the journal reference for the original sequence. The Medline link will give you an abstract etc. You will probably find that:

The PubMed server at http://www.ncbi.nlm.nih.gov/Entrez/ is a far better tool for browsing Medline that what is offered with SRS. Especially powerful is its facility for finding [Related entries].

Additional questions:

"Effective researchers know how to find things out"

1. Who submitted the serum amyloid A (SAA) gene sequence for Canis familiaris?
2. What prosite motif defines the recA family of prokaryotic proteins? Which Dublin-based phylogeneticists used multiple-sequence alignment to define this motif?
3. What are the first and last 5 bases in the intron of the yeast actin gene with EMBL accession number V01288?
4. What is the map position of one of the human SAA genes (SwissProt: P02735)? What cross-reference database is most likely to have map position?
5. What mutation at what position causes phenylketonuria (PKU)? (hint: EMBL K03020) but then try SwissProt: P00439.
6. What bases define the ribosome binding site of the Bacteroides fragilis glnA gene? Perhaps start from the E.coli homolog SwissProt: P06711.
7. Why is the name Saarinen associated with life-threatening cardiac arrythmias? (Hint: not because of architectural flaws...try voltage gated potassium channels)
8. Are there more publicly available DNA sequences from Rodents or Prokaryotes? What about protein sequences?
9. Get a sample of mammalian introns. See what common features they have? Think how these common features might help splicing out the introns.

Entrez - http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is the US equivalent of SRS and is available from the NCBI webpage. You will most likely be familiar with Entrez for interrogating Medline, but the same engine can be pointed at DNA and protein databases. It is handy if you are familiar with the Entrez system and you want a sequence whose name or accession number you already know. At the top of the Entrez page change the Search [__] choice box from PubMed to the appropriate sort of database – the available options are listed on the Entrez page. If you want the sequence alone – to paste into some analysis page – change the Display [__] choice box to FASTA then click on [Save] or [Display] depending on whether you want a permanent or transitory copy of you proteins. Entrez has a more complex syntax for less straightforward queries.

Nucleic Acid Sequence Analysis

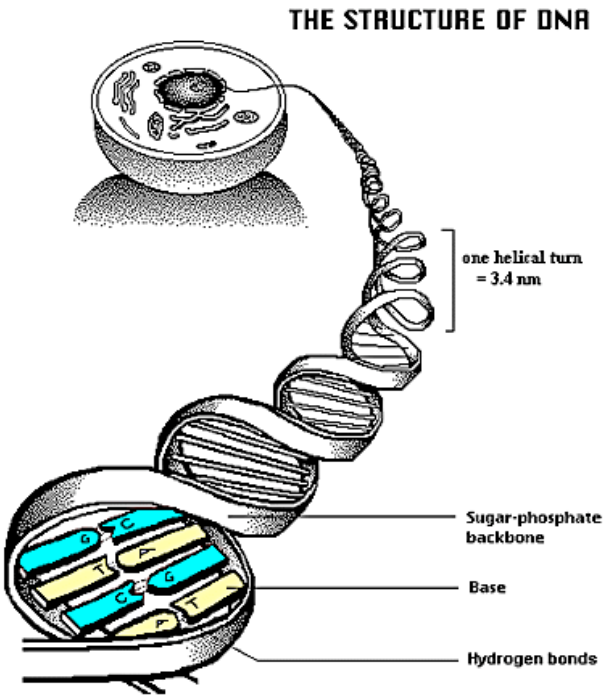
TOPICS:

1. Nucleic acids and the genetic code
2. Translating DNA in 6 frames.
3. Reverse complement & other tools.
4. Calculating some properties of DNA/RNA sequences.
5. Primer design.

1) Nucleic acids and the genetic code

Nucleic acids may be in the form of Deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) molecules containing the genetic information important for all cellular functions and heredity.

DNA is a long polymer of nucleotides to code for the sequence of amino acid during protein synthesis. DNA is said to carry the genetic 'blueprint' since it contains the instructions or information (called genes) needed to construct cellular components like proteins and RNA molecules.

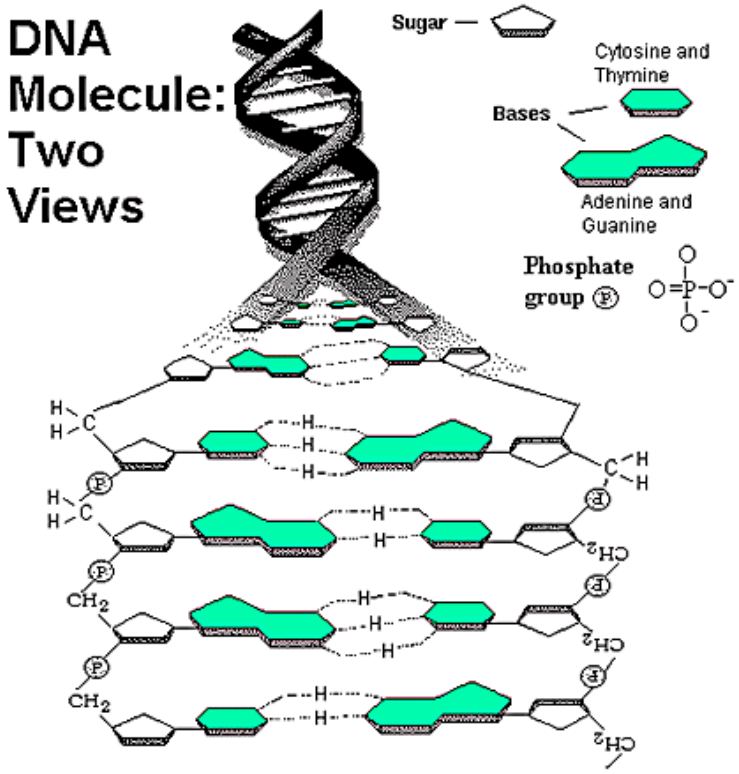


DNA is composed of two strands that twist together to form a helix. Each strand consists of alternating nucleotides. Each nucleotide consists of a phosphate (PO4) and pentose sugar (2-deoxyribose), and attached on the sugar is a nitrogenous base, which can be

adenine, thymine, guanine, or cytosine. The four nucleotides are given one letter abbreviations as shorthand for the four bases.

* A is for adenine
* G is for guanine
* C is for cytosine
* T is for thymine

See Appendix1 for more details.



Hence, DNA is a ladder-like helical structure. The two DNA strands are joined together at the center by pairing bases lined up with one another. Adenine pairs with thymine and guanine with cytosine. A and T are connected by two hydrogen bonds. G and C are connected by three hydrogen bonds. DNA is often described structurally as a twisting ladder. In this ladder, the “rungs” are the pairs of bases linked together, and the “sides” are the two separate sugar and phosphate backbones.

The double helix is important because it preserves all of the information-carrying features of a single DNA strand while at the same time introducing elements that make it easier for living cells to make copies of their DNA. Because every base pair in the double helix must match its pairing partner (A with T, C with G), we can easily determine the sequence of an unknown strand of DNA if its matching strand is known. For example, if one strand of a double helix has the nucleotide sequence

```
GATTCGTACG
```

then its complementary strand will be

```
CTAAGCATGC
```

forming a double helix

```
GATTCGTACG
||||||||||
CTAAGCATGC
```

2) Translating DNA in 6-frames:

Why six frames?

DNA code for amino acids using a Three-Letter genetic code. (See Appendix II for the complete genetic code.) Since we do not know where to start reading a DNA sequence, we need to look at six different options.

For example the sequence:

```
GATTCGTACG
||||||||||
CTAAGCATGC
```

Can be translated into six different amino acid strings. Looking at each strand separately:

GATTCGTACG

1 GAT Asp | TCG Ser | TAC Tyr | G

2 G | ATT Ile | CGT Arg | ACG Thr

3 GA | TTC Phe | GTA Val | CG

CTAAGCATGC

4 CTA Leu | AGC Ser | ATG Met | C

5 C | TAA # | GCA Ala | TGC Arg

6 CT | AAG Lys | CAT His | GC

Translate tool - http://www.expasy.ch/tools/dna.html

This tool allows the 6-frame translation of a nucleotide (DNA/RNA) sequence to a protein sequence in order to locate open reading frames in your sequence.

- Go to URL above.
- You can use the following phosphoglycerate kinase gene sequence from *Trypanosoma brucei* below or select from `phospho_kinase.txt`:

```
>Tb927.1.700 phosphoglycerate kinase Trypanosoma brucei
ATGACCCTTA ACGAGAAGAA GAGCATTAAT GAATGCGATC TTAAGGGAAA GAAGGTTCTT
ATCCGTGTTG ACTTTAATGT TCCCGTGAAA AACGGTAAGA TCACCAACGA CTACCGAATC
CGATCAGCTC TGCCAACGCT CAAGAAGGTT CTCACAGAAG GCGGCAGTTG TGTTCTCATG
AGCCACCTCG GGAGGCCGAA AGGTATTCCC ATGGCGCAAG CTGACAAAAT ACGGAGCACT
GGCGGTGTTC CCGGGTTCCA ACAGAAGGCA ACACTCAAAC CGGTAGCCAA GCGCCTCAGC
GAACTGCTAT TGAGGCCCGT CACATTCGCA CCTGACTGCC TGAATGCTGC AGATGTCGTC
TCTAAGATGT CTCCGGGCGA TGTTGTTCTG CTTGAAAATG TACGCTTTTA CAAAGAAGAG
GGCAGCAAGA AGGCAAAAGA ACGTGAAGCC ATGGCCAAGA TCCTTGCGTC ATATGGTGAT
GTTTACATCA GTGATGCTTT TGGTACAGCT CACCGTGACA GTGCTACCAT GACCGGAATT
CCAAAGATTT TGGGCAACGG TGCTGCCGGT TATTTGATGG AGAAGGAGAT TTCATACTTC
GCTAAGGTAC TTGGTAACCC GCCGCGTCCG CTGGTTGCTA TCGTTGGTGG AGCGAAAGTG
AGCGACAAGA TCCAACTTCT GGATAACATG TTGCAGCGCA TCGATTATCT CTTAATTGGT
GGTGCAATGG CATACACATT TCTGAAGGCT CAGGGTTACA GCATTGGAAA ATCGAAGTGC
GAGGAAAGTA AACTTGAATT TGCTCGATCC CTGCTGAAGA AGGCGGAGGA CCGCAAGGTG
CAGGTTATTC TTCCAATTGA TCATGTTTGC CACACGGAAT TCAAAGCTGT GGATTCTCCA
TTGATAACTG AGGATCAAAA CATCCCTGAA GGACATATGG CTCTGGATAT TGGTCCCAAG
ACTATTGAAA AATATGTTCA GACGATTGGG AAGTGTAAGA GCGCCATTTG GAACGGTCCC
ATGGGTGTAT TTGAAATGGT TCCTTATTCC AAAGGTACAT TTGCAATTGC GAAAGCCATG
GGTCGAGGAA CTCACGAGCA TGGACTCATG AGTATCATCG GTGGTGGTGA CAGCGCAAGT
GCAGCTGAGT TGAGCGGTGA GGCGAAGCGC ATGTCTCATG TTTCAACTGG TGGTGGTGCG
TCTTTGGAAC TCCTCGAGGG CAAAACGCTT CCCGGCGTTG CAGTATTGGA CGAAAAGTCG
GCGGTTGTGT CGTATGCCTC TGCAGGTACT GGAACTCTTT CTAACCGGTG GAGCTCTCTT
TAA
```

- Paste your sequence in the box provided & click "TRANSLATE SEQUENCE".
- You can choose 3 options
  - Verbose – puts Met & Stop to highlight start & stop codons.
  - Compact – useful if you want to use output in other programs.
  - Includes nucleotide sequence – nucleotide sequence is above the translation.
- This returns a 6-frame translation of your sequence. You can then choose the correct frame.

EMBOSS

transeq
Translate nucleic acid sequences

3) Reverse Complement & other tools:

There are many cases where you might want to obtain the reverse complement of a DNA sequence, for example the reverse complement is needed as a negative control when doing a DNA hybridisation experiment.

SEquence analysis using WEb Resources SeWeR

http://www.bioinformatics.org/SeWeR/

http://www.bioinformatics.org/SeWeR/

SeWeR is an integrated portal to common web-based services in bioinformatics. It has a large number of tools available online.

Nucleic Acid

- **Entrez** - Retrieve a DNA sequence from Genbank at NCBI server. Input type key-words
- **Webcutter -** One of the best programs for restriction analysis. Input type DNA.
- **Translate** - Translate your nucleotide (DNA/RNA) sequence to a protein sequence on Expasy server. Input type - DNA/RNA sequence.
- **GeneMark** - Predict ORF in your sequence. Input type: DNA sequence.

Protein

- **Entrez** - Retrieve a protein sequence from Genbank at NCBI server. Input type key-words
- **ProtParam -** Calculate different physico-chemical parameters of a protein sequence. Input type-protein sequence.

- **PSIPRED -** Secondary structure Prediction. Input type-protein sequence.
- **ScanProsite -** Search for PROSITE pattern in your protein sequence. Input type-protein sequence.

Database

- **BLASTN -** Search a nucleotide sequence against GenBank on NCBI server. Query type: DNA sequence.
- **BLASTP -** Search a protein sequence against the protein sequences on NCBI server. Query type-Protein Sequence.
- **Blocks Searcher -** Search BLOCKS database for similarity. Input type- protein or DNA sequence.
- **PUBMED -** Search bibliographic database at NCBI. Input type-key-word.

PCR

- **Primer3 -** Create PCR primers and Hybridization oligos from a DNA sequence. Input, DNA sequence.
- **CODEHOP -** Pick primers from multiple alingnment of protein sequences. Input type, BLOCKS.

Alignment

- **ClustalW -** Align multiple sequence. Input type, DNA or protein sequences.
- **Block Maker -** Finds conserved blocks in a group of two or more unaligned protein sequences. At least two protein sequences must be provided to make blocks. Each sequence must have a unique name of 10 characters or less. All sequences must be of same format (FASTA). Input type- protein sequences.

Tools

- **ReadSeq -** It automatically recognizes the input sequence type and convert it into a format of choice. Input type-DNA/protein sequence(s)of different formats.
- **CAP -** Contig assembly program (CAP). Input type-DNA sequences in FASTA format.
- **Clean/Inverse complement -** You can **Inverse-complement** the sequence or Clean the sequence. In either case SeWeR will filter out only A/T/G/C/N from the query. All spaces, numbers, line-breaks will be removed from the sequence

EMBOSS

revseq
Reverse and complement a sequence

eprimer3
Picks PCR primers and hybridization oligos

primersearch
Searches DNA sequences for matches with primer pairs

restrict
Finds restriction enzyme cleavage sites

transeq
Translate nucleic acid sequences

prettyseq
Output sequence with translated ranges

plotorf
Plot potential open reading frames

showorf
Pretty output of DNA translations

splitter
Split a sequence into (overlapping) smaller sequences

Exercise: Paste in the phosphoglycerate kinase gene sequence from *Trypanosoma brucei* for each application. Pay particular attention to the options available: these will give you clues about standard practice.

See if you can repeat the exercise using the EMBOSS program's.

See Appendix1 and Appendix2 for details about the genetic code.

4) Oligo Calculator - http://mbcf.dfci.harvard.edu/docs/oligocalc.html

Tool to calculate the **length**, **%GC content**, **Melting temperature (Tm)** the midpoint of the temperature range at which the nucleic acid strands separate, **Molecular weight**, & what an **OD** = 1 is in picoMolar of your input nucleic acid sequence.

Many of these parameters are useful in primer design (see next section) and in other areas of molecular biology.

- Go to URL above.

- Paste the phosphoglycerate kinase gene sequence from *Trypanosoma brucei* in the box provided and click "**Calculate**".

Example:

>Tb927.1.700 phosphoglycerate kinase Trypanosoma brucei
Length = 1323
% GC content = 49
Tm = 84 °C
Molecular Weight = 409839 daltons (g/M)
OD of 1 = 69 picoMolar

EMBOSS

dan
Calculates DNA RNA/DNA melting temperature

eprimer3
Picks PCR primers and hybridization oligos

Protein Sequence Analysis

TOPICS

- Physico-chemical properties.
- Cellular localization.
- Signal peptides.
- Transmembrane domains.
- Post-translational modifications.
- Motifs & domains.
- Secondary structure.
- Other resources.

ExPASy - http://www.expasy.ch/

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) protein and proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures. Besides the tools that we will introduce in this manual there are many other applications available at this website that you should take some time to have a look at.

1) Physico-chemical properties:

ProtParam tool – http://www.expasy.ch/tools/protparam.html

Or use

http://www.bioinformatics.org/SeWeR/

Calculates lots of physico-chemical parameters of a protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)

Example: Human BRCA 1

You can paste the gene sequence brca1 from the course website.

- At ExPASy → "Proteomics and sequence analysis tools" → "Primary structure analysis".

- Click on the "ProtParam" link.
- Paste your sequence in the box provided
- The sequence must be written using the one letter amino acid code:
- Press the "Compute parameters" button.
- The output for this sequence is shown below.

Number of amino acids: 1863
Molecular weight: 207720.8
Theoretical pI: 5.29
Amino acid composition:
Ala (A) 84 4.5%
Arg (R) 76 4.1%
Etc etc
Thr (T) 111 6.0%
Trp (W) 10 0.5%
Tyr (Y) 31 1.7%
Val (V) 101 5.4%
Asx (B) 0 0.0%
Glx (Z) 0 0.0%
Xaa (X) 0 0.0%

Total number of negatively charged residues (Asp + Glu): 283
Total number of positively charged residues (Arg + Lys): 213
Atomic composition:
Carbon C 8908
Hydrogen H 14246
Nitrogen N 2554
Oxygen O 3014
Sulfur S 74
Formula: $C_{8908}H_{14246}N_{2554}O_{3014}S_{74}$
Total number of atoms: 28796
Extinction coefficients:
Conditions: 6.0 M guanidium hydrochloride
0.02 M phosphate buffer
pH 6.5
Extinction coefficients are in units of M^{-1} cm^{-1}.
The first table lists values computed assuming ALL Cys residues appear as half cystines, whereas the second table assumes that NONE do.

```
276 278 279 280 282
nm nm nm nm nm
Ext. coefficient 102140 102194 100935 99220 95840
Abs 0.1% (=1 g/l) 0.492 0.492 0.486 0.478 0.461
276 278 279 280 282
nm nm nm nm nm
Ext. coefficient 98950 99400 98295 96580 93200
Abs 0.1% (=1 g/l) 0.476 0.479 0.473 0.465 0.449
```

Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).
Instability index:
The instability index (II) is computed to be 54.68

This classifies the protein as unstable.
Aliphatic index: 69.01
Grand average of hydropathicity (GRAVY): -0.785

EMBOSS

pepinfo
Plots simple amino acid properties

pepstats
Protein statistics

charge
Protein charge plot

iep
Calculates the isoelectric point of a protein

2) Cellular localization:

PSORT - http://psort.nibb.ac.jp/form2.html

PSORT, a program to predict the subcellular localization sites of proteins from their amino acid sequences. This program makes use of the fact that proteins destined for particular subcellular localizations have distinct amino acid properties particularly in their N-terminal regions. These properties can be used to predict whether a protein is localized in the cytoplasm, nucleus, mitochondria, or is retained in the ER, or destined for the lysosome (vacuolar) or the peroxisome. There is a detailed page of output that we can probably ignore. At the end of the output the percentage likelihood of the subcellular localization is given. If you want to learn more about the output and how subcellular localization is determined please see the user manual at:

http://psort.nibb.ac.jp/helpwww2.html

Example: Human ETS-1 protein.

You can paste the gene sequence ets-1 from the course website.

- At http://psort.nibb.ac.jp/form2.html
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Press the submit button.
- The output for this sequence is shown below.

There are a number parameters measured by this program which you can read about as links from the output file. By scrolling to the bottom of the output you can see the probability that this sequence is nuclear, cytoplasmic, peroxisomal, vacuolar or cytoskeletal. PSORT predicts that ETS-1 is nuclear with a high probability. The fact that ETS-1 is localized in the nucleus has been previously experimentally determined.

Results of Subprograms

```
PSG: a new signal peptide prediction method
N-region: length 8; pos.chg 2; neg.chg 1
H-region: length 6; peak value 1.89
PSG score: -2.51
```

Results of the k-NN Prediction

```
k = 9/23
73.9 %: nuclear
13.0 %: cytoplasmic
4.3 %: peroxisomal
4.3 %: vacuolar
4.3 %: cytoskeletal
>> prediction for QUERY is nuc (k=23)
```

3) Signal peptides:

Proteins destined for secretion, operation with the endoplasmic reticulum, lysosomes and many transmembrane proteins are synthesized with leading (N-terminal) 13 – 36 residue signal peptides.

SignalP - http://www.cbs.dtu.dk/services/SignalP/

The SignalP WWW server can be used to predict the presence and location of signal peptide cleavage sites in your proteins. It can be useful to know whether your protein has a signal peptide as it indicates that it may be secreted from the cell. Furthermore, proteins in their active form will have their signal peptides removed, if you can determine the length of the signal peptide then you can calculate the size of the protein minus the signal peptide.

Example: Human Beta-defensin; sp|Q09753|BD01_HUMAN

You can paste the gene sequence HBD1 from the course website.

At ExPASy → "Post-translational modification prediction".

Click on the "SignalP" link.

Paste your sequence in the box provided

The sequence must be written using the one letter amino acid code:

It is recommend that the N-terminal part only (not more than 50-70 amino acids) of the sequences is submitted. A longer sequence will increase the risk of false positives and make the graphical output difficult to read. The new version now automatically truncates input sequences.

Choose one or more group of organisms for the prediction by clicking the check-box next to the group(s):

If no groups are indicated, predictions from all three groups will be returned.

A graphical output (in Postscript format) of the prediction will be available, if the "Include graphics"-button is checked.

Press the "Submit sequence" button.

A WWW page will return the results when the prediction is ready. Response time depends on system load. The output for this sequence is shown below

C score = raw cleavage site score

The output score from networks trained to recognize cleavage sites vs. other sequence positions. Trained to be: High at position +1 after the cleavage site and low at all other positions.

S score = signal peptide score

The output score from networks trained to recognize signal peptide vs. non-signal-peptide positions. Trained to be: High at position before the cleavage site and low at all other positions.

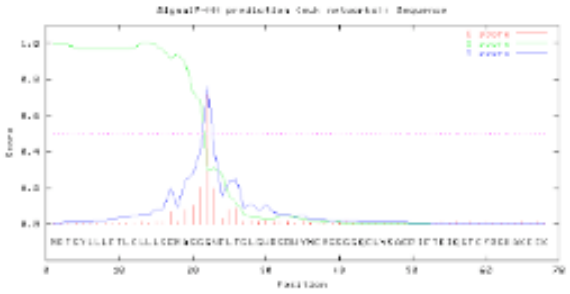
Y score = combined cleavage site score

The prediction of cleavage site location is optimized by observing where the C-score is high and the S-score changes from a high to a low value.

For each sequence, SignalP will report the maximal C, S, and Y scores, and the mean S-score between the N-terminal and the predicted cleavage site. These values are used to distinguish between signal peptides and non-signal peptides. If your sequence is predicted to have a signal peptide, the cleavage site is predicted to be immediately before the position with the maximal Y-score.

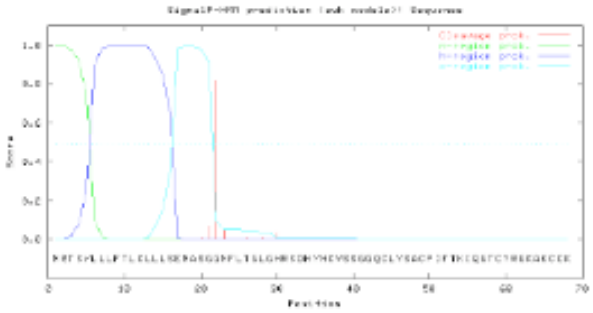
The Human beta-defensin protein has a predicted signal peptide from position 1 to 21 and a potential cleavage site exists between positions 21 and 22. These predictions correspond exactly to the SWISS-PROT annotation for this protein (accession Q09753).

SignalP-NN result:



```
# data
>Sequence length = 68
# Measure Position Value Cutoff signal peptide?
max. C 22 0.710 0.32 YES
max. Y 22 0.761 0.33 YES
max. S 14 0.998 0.87 YES
mean S 1-21 0.943 0.48 YES
D 1-21 0.852 0.43 YES
# Most likely cleavage site between pos. 21 and 22: ASG-GN
```

SignalP-HMM result:



```
# data
>Sequence
Prediction: Signal peptide
Signal peptide probability: 1.000
Signal anchor probability: 0.000
Max cleavage site probability: 0.818 between pos. 21 and 22
```

EMBOSS

sigcleave
Reports protein signal cleavage sites

4) Transmembrane domains:

Tmpred - http://www.ch.embnet.org/software/TMPRED_form.html

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. The presence of transmembrane domains is an indication that the protein is located on the cell surface.

Example: Human chemokine receptor 4 protein sequence NP_003458.1

You can paste the gene sequence chemo4 from the course website.

At ExPASy → "Topology prediction".

Click on the link to Tmpred.

Paste your sequence in the box provided in one of the supported formats e.g.

plain text, SwissProt_ID or AC, etc.

You may change the minimal and maximal length of the hydrophic part of the transmembrane helix but unless you have reason to do so you should accept the defaults i.e. 17 and 33. ~22 residues is the same length as the width of a lipid bilayer.

Click the "Run Tmpred" button to start the search.

The output is given in 3 parts 1, 2 and 3 (see below).

Part 1: lists all the significant predictions of possible transmembrane helices

in this case there are 7 helices predicted but at this stage we do not know the orientation of the helices so there are 2 tables, the first with the helices orientated from the inside to the outside and vice versa for the second.

Part 2: shows which inside->outside helices correspond to the outside -> inside helices and indicates which orientation is most likely.

Part 3: proposes the strongly preferred model for the transmembrane domain structure of the protein and also an alternative model.

A graphic of the prediction is also available (not shown here)

These predictions correspond well but not exactly to the SWISS-PROT annotation for this protein (accession P30991).

Tmpred output

Sequence: MEG...HSS, length: 352

Prediction parameters: TM-helix length between 17 and 33

1. Possible transmembrane helices

The sequence positions in brackets denominate the core region. Only scores above 500 are considered significant.

```
Inside to outside helices : 7 found
from to score center
39 ( 46) 62 ( 62) 1962 54
78 ( 85) 105 ( 103) 1623 95
114 ( 114) 133 ( 130) 1352 122
155 ( 157) 175 ( 173) 1716 165
204 ( 206) 223 ( 223) 2052 214
240 ( 240) 261 ( 259) 2840 251
286 ( 286) 305 ( 305) 1241 295
Outside to inside helices : 7 found
from to score center
47 ( 47) 63 ( 63) 2568 55
78 ( 78) 96 ( 96) 1331 86
111 ( 114) 132 ( 132) 1740 122
```

```
155 ( 157) 173 ( 173) 1197 165
204 ( 204) 223 ( 223) 2404 214
240 ( 242) 259 ( 259) 2037 251
283 ( 286) 305 ( 305) 1703 294
```

---

2. Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.
Helices shown in brackets are considered insignificant. A "+"-symbol indicates a preference of this orientation. A "++"-symbol indicates a strong preference of this orientation.

```
Inside->outside | outside->inside
39- 62 (24) 1962 | 47- 63 (17) 2568 ++
78- 105 (28) 1623 ++ | 78- 96 (19) 1331
114- 133 (20) 1352 | 111- 132 (22) 1740 ++
155- 175 (21) 1716 ++ | 155- 173 (19) 1197
204- 223 (20) 2052 | 204- 223 (20) 2404 ++
240- 261 (22) 2840 ++ | 240- 259 (20) 2037
286- 305 (20) 1241 | 283- 305 (23) 1703 ++
```

---

3. Suggested models for transmembrane topology

These suggestions are purely speculative and should be used with extreme caution since they are based on the assumption that all transmembrane helices have been found. In most cases, the Correspondence Table shown above or the prediction plot that is also created should be used for the topology assignment of unknown proteins.

2 possible models considered, only significant TM-segments used

```
---→ STRONGLY preferred model: N-terminus outside
7 strong transmembrane helices, total score : 14594
# from to length score orientation
1 47 63 (17) 2568 o-I
2 78 105 (28) 1623 I-o
3 111 132 (22) 1740 o-I
4 155 175 (21) 1716 I-o
5 204 223 (20) 2404 o-I
6 240 261 (22) 2840 I-o
7 283 305 (23) 1703 o-I
----→ alternative model
7 strong transmembrane helices, total score : 11172
# from to length score orientation
1 39 62 (24) 1962 I-o
2 78 96 (19) 1331 o-I
3 114 133 (20) 1352 I-o
4 155 173 (19) 1197 o-I
5 204 223 (20) 2052 I-o
6 240 259 (20) 2037 o-I
7 286 305 (20) 1241 I-o
```

EMBOSS

tmap
Displays membrane spanning regions

5) Post-translational modifications:

After translation has occurred proteins may undergo a number of posttranslational modifications. These can include the cleavage of the pro- region to release the active protein, the removal of the signal peptide and numerous covalent modifications such as, acetylations, glycosylations, hydroxylations, methylations and phosphorylations. Posttranslational modifications such as these may alter the molecular weight of your protein and thus its position on a gel. There are many programs available for predicting the presence of posttranslational modifications, we will take a look at one for the prediction of type O-glycosylation sites in mammalian proteins. Remember these programs work by looking for consensus sites and just because a site is found does not mean that a modification definitely occurs.

NetOGlyc - http://www.cbs.dtu.dk/services/NetOGlyc/

Prediction of type O-glycosylation sites in mammalian proteins. This program works by comparing the input sequence to a database of known and verified mucin type O-glycosylation sites extracted from O-GLYCBASE.

Example: Human CD1D sp|P15813|CD1D_HUMAN

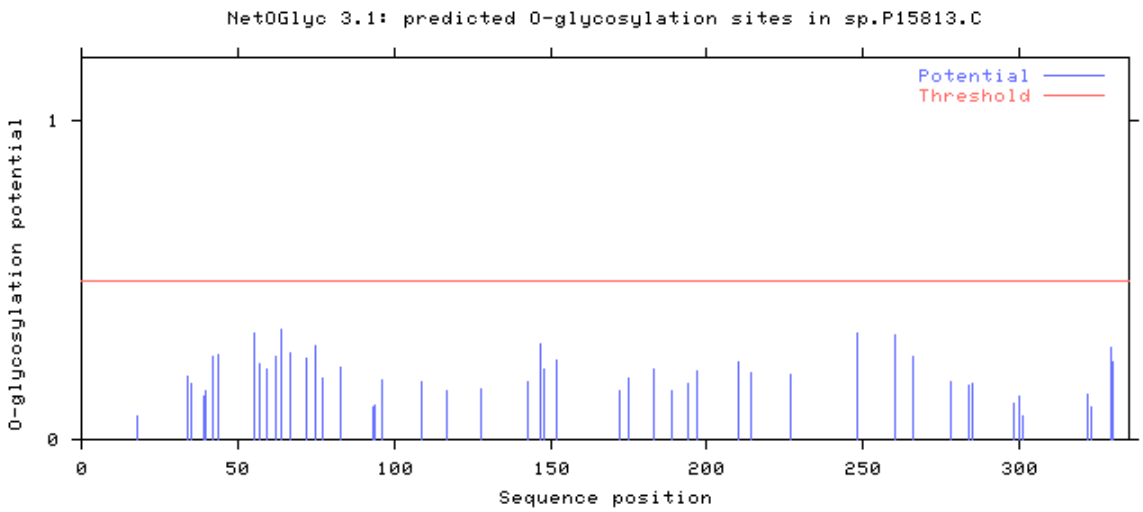
You can paste the gene sequence cd1d from the course website.

- At ExPASy → "Post-translational modification".
- Click on the link to "NetOGlyc".
- Paste your sequence in the box provided in FASTA format.
- Check "generate graphics" and click the submit button.
- The output for this program is shown below (graphics not shown).
- This program predicts potential O-glycosylation sites at Threonine 64 and Serine 214.

NetOGlyc 3.1 Prediction Results

```
Name:  sp_P15813_C            Length:  335
MGCLLFLLLWALLQAWGSAEVPQRLFPLRCLQISSFANSSWTRTDGLAWLGELQTHSWSNDSDTVRSLKPW
SQGTFSDQQWETLQHIFRVYRSSFTRDVKEFAKMLRLSYPLELQVSAGCEVHPGNASNNFFHVAFQGKDIL
SFQGTSWEPTQEAPLWVNLAIQVLNQDKWTRETVQWLLNGTCPQFVSGLLESGKSELKKQVKPKAWLSRGP
SPGPGRLLLVCHVSGFYPKPVWVKWMRGEQEQQGTQPGDILPNADETWYLRATLDVVAGEAAGLSCRVKHS
SLEGQDIVLYWGGSYTSMGLIALAVLACLLFLLIVG FTSRFKRQTSYQGVL
_________________.....................................................
.............................................................................
.............................................................................
.............................................................................
.................................. ..............
Name                        S/T   Pos  G-score I-score Y/N  Comment ---
-----------------------------------------------------------------------
sp_P15813_C                   S     18   0.075   0.079   .   -
sp_P15813_C                   S     34   0.198   0.051   .   -
sp_P15813_C                   S     35   0.177   0.037   .   -
```

```
sp_P15813_C                     S     39   0.133   0.022   .   -
sp_P15813_C                     S     40   0.153   0.023   .   -
sp_P15813_C                     T     42   0.259   0.054   .   -
sp_P15813_C                     T     44   0.267   0.055   .   -
sp_P15813_C                     T     55   0.334   0.052   .   -
sp_P15813_C                     S     57   0.239   0.036   .   -
sp_P15813_C                     S     59   0.221   0.066   .   -
sp_P15813_C                     S     62   0.261   0.032   .   -
sp_P15813_C                     T     64   0.347   0.033   .   -
sp_P15813_C                     S     67   0.269   0.071   .   -
sp_P15813_C                     S     72   0.256   0.076   .   -
sp_P15813_C                     T     75   0.292   0.024   .   -
sp_P15813_C                     S     77   0.193   0.047   .   -
sp_P15813_C                     T     83   0.226   0.041   .   -
sp_P15813_C                     S     93   0.099   0.051   .   -
sp_P15813_C                     S     94   0.109   0.020   .   -
sp_P15813_C                     T     96   0.185   0.078   .   -
sp_P15813_C                     S    109   0.180   0.069   .   -
sp_P15813_C                     S    117   0.153   0.092   .   -
sp_P15813_C                     S    128   0.156   0.050   .   -
sp_P15813_C                     S    143   0.183   0.058   .   -
sp_P15813_C                     T    147   0.300   0.018   .   -
sp_P15813_C                     S    148   0.222   0.030   .   -
sp_P15813_C                     T    152   0.248   0.047   .   -
sp_P15813_C                     T    172   0.153   0.081   .   -
sp_P15813_C                     T    175   0.195   0.021   .   -
sp_P15813_C                     T    183   0.222   0.021   .   -
sp_P15813_C                     S    189   0.155   0.079   .   -
sp_P15813_C                     S    194   0.177   0.053   .   -
sp_P15813_C                     S    197   0.217   0.024   .   -
sp_P15813_C                     S    210   0.242   0.091   .   -
sp_P15813_C                     S    214   0.210   0.333   .   -
sp_P15813_C                     S    227   0.204   0.032   .   -
sp_P15813_C                     T    248   0.334   0.273   .   -
sp_P15813_C                     T    260   0.327   0.033   .   -
sp_P15813_C                     T    266   0.261   0.042   .   -
sp_P15813_C                     S    278   0.183   0.053   .   -
sp_P15813_C                     S    284   0.170   0.025   .   -
sp_P15813_C                     S    285   0.175   0.027   .   -
sp_P15813_C                     S    298   0.111   0.017   .   -
sp_P15813_C                     T    300   0.134   0.037   .   -
sp_P15813_C                     S    301   0.073   0.080   .   -
sp_P15813_C                     T    322   0.141   0.070   .   -
sp_P15813_C                     S    323   0.103   0.031   .   -
sp_P15813_C                     T    329   0.290   0.030   .   -
sp_P15813_C                     S    330   0.245   0.026   .   - --------
------------------------------------------------------------------------
```



6) Motifs and Domains

If you want to determine the function of a protein the first tool of choice is homology searching (see day 4). Unless this finds you a match with a well characterized protein comprehending the entire length of yours you should look for motifs and domains in your protein. To determine if your protein sequence contains known motifs or conserved domain structures you should search the protein against one of the motif or profile databases. There are many of these available but we will discuss ProfileScan (now called myHits), which allows you to search both the Prosite and Pfam databases simultaneously. See the documentation for more details.

ProfileScan - http://hits.isb-sib.ch/cgi-bin/PFSCAN

Example: Human CFTR sp|P13569|CFTR_HUMAN

You can paste the gene sequence cftr from the course website.

- Go to the URL above
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Tick the motif databases you wish to search, other parameters should be OK.
- Press the “scan” button.

The output for this program is too large to show here, but it gives lots of detail about motifs in the CFTR protein identifying potential: ABC transporters family signature; ATP/GTP-binding site motif A (P-loop); Protein kinase C phosphorylation sites; N-glycosylation sites; Casein kinase II phosphorylation site; N-myristoylation sites; cAMP- and cGMP-dependent protein kinase phosphorylation site; Bipartite nuclear localization signal; NACHT-NTPase domain profile; Guanylate kinase domain profile etc.

Remember that these programs only tell you are that there is a motif present and thus there is the potential for these modifications and functions to occur. It is up to you to determine experimentally which are real but at least you now know what to look for.

7) Secondary Structure Prediction

If protein structure, even secondary structure, can be accurately predicted from the now abundantly available gene and protein sequences, such sequences become immensely more valuable for the understanding of drug-design, the genetic basis of disease, the role of protein structure in its enzymatic, structural, and signal transduction functions, and basic physiology from molecular to cellular, to fully systemic levels. In short, the solution of the protein structure prediction problem (and the related protein folding problem) will bring on the second phase of the molecular biology revolution (Munson et al., 1994).

JPRED – http://www.compbio.dundee.ac.uk/www-jpred/

Jpred is an Internet web server that takes either a protein sequence or a multiple alignment of protein sequences, and predicts secondary structure. It works by combining a number of modern, high quality prediction methods to form a consensus. Please be aware that secondary structure prediction is an extremely complex problem that is under intensive research and we are still at a relatively primitive stage. We cannot discuss the details of protein secondary structure here but if you are interested in this area we recommend that you take a look at any major biochemistry textbook. Essentially protein secondary structure consists of 3 major conformations; the α Helix, the β pleated sheet and the coil conformation.

Example: Human alpha 1 hemoglobin. NP_000549.1

You can paste the gene sequence hbb from the course website.

- Go to the website
- Paste your sequence in the box provided.
- The defaults are OK.
- Click “Makepredictions!”
- If your sequence already has had its structure predicted or experimentally determined it will be in here and you can follow the link to PDB for information on the structure of your protein.
- If your protein is in PDB you can view your protein secondary structure using RasMol (To download RasMol see the course website for a link).
- Once you have RasMol running you can open your structure in it a view it using a number of different options.

Otherwise continue with prediction

- The program may take a long time so you can save a bookmark and return to your results later or choose to have your results e-mailed to you.

- There are a number of options to view the output, view your output in HTML format (option 4).
- The complete output is too large to show here (see webpage).
- Scroll down through the output until you get to "Jpred" output. The line of output beside this is the consensus secondary structure for your sequence. H= Helices E= strands C= coils.

Printed sources about Bioinformatics and the Internet.

Briefings in Bioinformatics - a journal aimed at users rather than developers with useful review and how-to articles.

Books:

Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins. Andreas Baxevanis & B.F.Francis Ouellette (Eds). John Wiley & Sons 2nd Ed 2001; ISBN: 0471-38390-2 The Course text book!

Fundamentals of Molecular Evolution. W-H Li and D Graur. Sinauer 1991. ISBN 0 87893 452 9

Fundamentals of Molecular Evolution. D Graur and W-H Li . Sinauer 2000. ISBN 0-87893-266-6

PAUP 4.0 Phylogenetic Analysis Using Parsimony (and other methods) Manual. David L Swofford. Sinauer 1999. 0 87893 801 X

Introduction to Bioinformatics. TK Attwood & DJ Parry-Smith. Addison Wesley Longman 1999. ISBN 0582 32788 1.

Molecular Evolution: a phylogenetic approach. RDM Page and EC Holmes. Blackwell 1998. ISBN: 0-86542-889-1

Bioinformatics for Dummies. Notredame and Claverie. 2003

Articles:

Baldauf, SL (2003) Phylogeny for the faint of heart: a tutorial. TIG 19(6): 345-351.

APPENDIX I

Nucleotide and Amino Acid Codes

Nucleotides

Description	Abbreviation
Adenosine	A
Thymidine	T
Cytosine	C
Guanosine	G
Uridine	U
Any nucleotide (A, T, C or G)	N
G or A	R
A or T	W
C or T	Y
A or C	M
G or T	K
G or C	S
Not G (A or C or T)	H
Not A (C or G or T)	B
Not T (A or C or G)	V
Not C (A or G or T)	D

Amino Acids

Full name	Single letter code	Three letter code	Codons
Alanine	A	Ala	4
Arginine	R	Arg	6
Asparagine	N	Asn	2
Aspartic acid	D	Asp	2
Cysteine	C	Cys	2
Glutamine	Q	Gln	2
Glutamic acid	E	Glu	2
Glycine	G	Gly	4
Histidine	H	His	2
Isoleucine	I	Ile	3
Leucine	L	Leu	6
Lysine	K	Lys	2
Methionine	M	Met	1
Phenylalanine	F	Phe	2
Proline	P	Pro	4
Serine	S	Ser	6
Threonine	T	Thr	4
Tryptophan	W	Trp	1

Tyrosine	Y	Tyr	2
Valine	V	Val	4

SEQUENCE SYMBOLS

Nucleotides

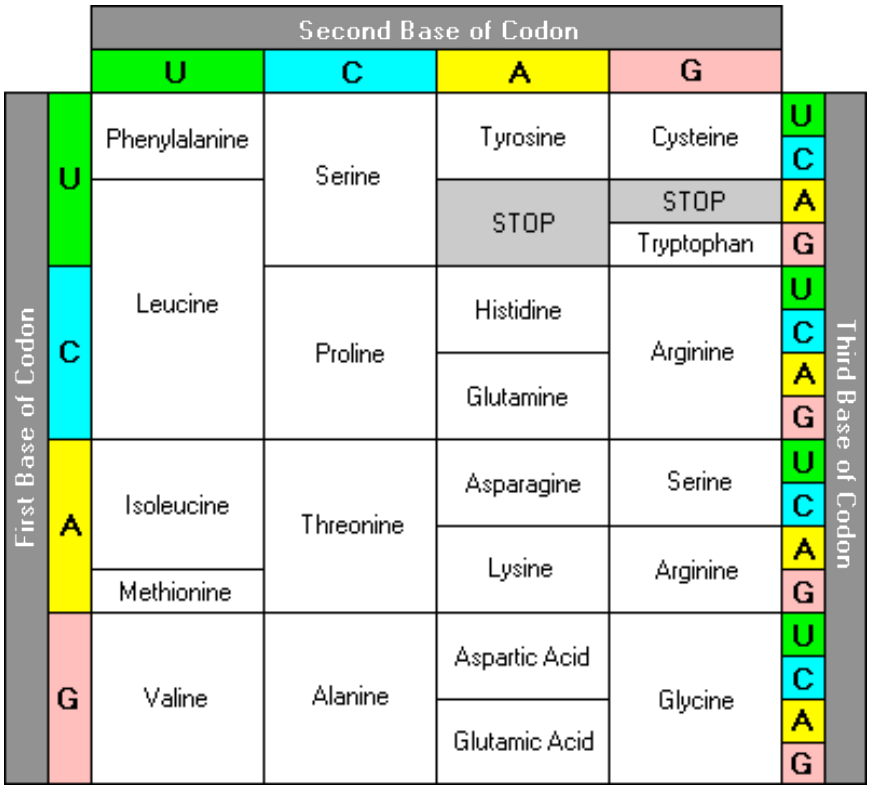
IUBcode	MEANING	COMPLEMENT
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X/N	G or A or T or C	X
.	not	G or A or T or C

Amino Acids

SYMBOL	MEANING	CODONS	IUB code
A	Ala	GCT, GCC, GCA, GCG	!GCX
B	Asp, Asn	GAT, GAC, AAT, AAC	!RAY
C	Cys	TGT, TGC	!TGY
D	Asp	GAT, GAC	!GAY
E	Glu	GAA, GAG	!GAR
F	Phe	TTT, TTC	!TTY
G	Gly	GGT, GGC, GGA, GGG	!GGX
H	His	CAT, CAC	!CAY
I	Ile	ATT, ATC, ATA	!ATH
K	Lys	AAA, AAG	!AAR
L	Leu	TTG, TTA, CTT, CTC, CTA, CTG	!TTR, CTX, YTR
M	Met	ATG	!ATG
N	Asn	AAT, AAC	!AAY
P	Pro	CCT, CCC, CCA, CCG	!CCX
Q	Gln	CAA, CAG	!CAR
R	Arg	CGT, CGC, CGA, CGG, AGA, AGG	!CGX, AGR, MGR
S	Ser	TCT, TCC, TCA, TCG, AGT, AGC	!TCX, AGY
T	Thr	ACT, ACC, ACA, ACG	!ACX
V	Val	GTT, GTC, GTA, GTG	!GTX
W	Trp	TGG	!TGG
X	Unknown		!XXX
Y	Tyr	TAT, TAC	!TAY
Z	Glu, Gln	GAA, GAG, CAA, CAG	!SAR
*	Terminator	TAA, TAG, TGA	!TAR, TRA

APPENDIX II

The Universal Genetic Code.




```
Phe   UUU  Ser UCU   Tyr UAU   Cys  UGU
      UUC       UCC       UAC       UGC
Leu   UUA       UCA  ter UAA   ter  UGA
      UUG       UCG  ter UAG   Trp  UGG
Leu   CUU  Pro CCU   His CAU   Arg CGU
      CUC       CCC       CAC       CGC
      CUA       CCA  Gln CAA       CGA
      CUG       CCG       CAG       CGG
Ile   AUU  Thr ACU   Asn AAU   Ser  AGU
      AUC       ACC       AAC       AGC
      AUA       ACA  Lys AAA   Arg  AGA
Met   AUG       ACG       AAG       AGG
Val   GUU  Ala GCU   Asp GAU   Gly  GGU
      GUC       GCC       GAC       GGC
      GUA       GCA  Glu GAA       GGA
      GUG       GCG       GAG       GGG
```

Exceptions to the Universal Code:

#1: Yeast Mitochondrial Code: CUN=T AUA=M UGA=W
#2: Mitochondrial Code of Vertebrates: AGR=* AUA=M UGA=W
#3: Mitochondrial Code of Filamentous fungi: UGA=W
#4: Mitochondrial Code of Insects and platyhelminths: AUA=M UGA=W AGR=S
#5: Nuclear Code of Candida cylindracea (see nature 341:164): CUG=S
#6: Nuclear Code of Ciliata: UAR = Q
#7: Nuclear Code of Euplotes: UGA=C
#8: Mitochondrial Code of Echinoderms: UGA=W AGR=S AAA=N
#9: Mitochondrial Code of Ascidaceae: UGA=W AGR=G AUA=M
#10: Mitochondrial Code of Platyhelminthes: UGA=W AGR=S UAA=Y AAA=N
#11: Nuclear Code of Blepharisma: UAG=Q